Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 1, February 2026, pp. 693~701
ISSN: 2302-9285, DOI: 10.11591/eei.v15i1.8035

Data mining approach for stunting clusters in Jumput Rejo

Amir Ali[1,2], Purwanto[3], Mundakir[4]
[1]Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia
[2]Medical Record and Health Information, Dr. Soetomo Hospital Foundation Health College, Surabaya, Indonesia
[3]Information Systems, Graduate School, Diponegoro University, Semarang, Indonesia
[4]Department of Nursing Science, Faculty of Health Sciences, Muhammadiyah Surabaya University, Surabaya, Indonesia

Article Info

Article history:

Received Dec 16, 2023
Revised Jul 22, 2025
Accepted Dec 6, 2025

Keywords:

Anthropometric
Cluster
K-Means
Prevalence stunting
Stunting

ABSTRACT

The target of reducing the stunting prevalence rate by 14% in 2024 which has been set by the government needs to be of concern to be implemented by the local health office. The purpose of the research is to cluster toddler anthropometry data with data mining algorithm. Optimize K-Means (KM) algorithm with elbow method use to cluster toddler anthropometry data (sex, height, weight, age, and health care center). A set of 580 children's anthropometric measurements were analyzed and categorized based on their similarity. Cluster 1 comprises 150 members and exhibits a narrower range of age and height values compared to the other clusters. Cluster 2, with 124 members, displays a broader range of age and height values compared to both Cluster 1 and Cluster 3. Cluster 3, consisting of 150 members, demonstrates age and height values that are higher than Cluster 1 but lower than Cluster 2 and Cluster 4. Finally, Cluster 4, encompassing 156 members, exhibits age and height values that are higher than those in the other clusters that many children are stunted based on standard anthropometric table for assessing children's nutritional status. The cluster optimization yielded four distinct clusters, which will serve as the input for identifying clusters during the data grouping process using the KM algorithm.

Corresponding Author:

Amir Ali
Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University
Imam Barjo Pleburan, Kecamatan Semarang Selatan, Semarang, Jawa Tengah, Indonesia
Email: amir_ali@stikes-yrsds.ac.id

1. INTRODUCTION

Improving children's growth to meet standard benchmarks is a key objective outlined in the sustainable development goals (SDGs), which require resolution by 2030. Specifically, the target within the SDGs aims to eradicate all forms of malnutrition, including stunting, as detailed in associated metadata documents. Recognized as crucial, this target underscores that chronic or recurring malnutrition can hinder children from reaching their full physical and cognitive capabilities [1]. Therefore it is necessary to strengthen and develop systems, data, information, research, and technological innovation, especially information technology. Policies and interventions should prioritize children with low birth weight, those aged over 30 months, and children born to underweight mothers. Research carried out by Bitew *et al*. [2] reveals significant regional disparities in childhood undernutrition and demonstrates the potential applicability of widely used machine learning algorithms in forecasting the factors contributing to child stunting, wasting, and underweight issues in Ethiopia. According to previous research, a variety of factors contribute to childhood malnutrition. There are several groups of factors that affect children's nutritional status, including: i) immediate factors like the child's age, gender, birth weight, inadequate food consumption, and infections; ii) remote factors such as socio-cultural, economic, environmental, and climatic

variables; and iii) intermediary factors like socioeconomic status, maternal ethnicity, and the educational level of both parents [3]. The process of growth restriction commonly initiates during pregnancy and persists throughout the initial two years following birth. The significant and permanent harm to both the body and cognitive abilities that accompanies impaired growth poses a significant obstacle to human progress. Recognizing the immense scale of stunted growth and its disastrous outcomes has led to its recognition as a crucial global health concern, receiving attention and prioritization at the highest international levels. In order to address this issue, global targets for reducing stunting have been established for 2025 and beyond [4]. UNICEF's conceptual framework on the causes of malnutrition is expanded upon by the WHO's framework on childhood stunting. In this updated version, both stunted growth and development are emphasized as central aspects, acknowledging their shared underlying causes, particularly during the critical period from 9 to 24 months. Approaches aimed at fostering and safeguarding wholesome growth are anticipated to positively impact children's physical, mental, socio-emotional, and intellectual progress and maturation [5]. Malnutrition is common and can negatively affect various clinical outcomes in oncology patients [6]. Millions of children worldwide suffer from childhood stunting, the most common form of malnutrition. In societies where being shorter than average is accepted as normal, stunting often goes unrecognized despite its widespread occurrence and established criteria for identification and assessment [4]. Reduced linear growth, a widespread issue in children from low- and middle-income nations, arises from fetal and/or early childhood exposure to insufficient nutrition and infectious ailments [7].

The clustering algorithm is a data mining approach. Sorting a collection of items into clusters according to their shared traits is the aim of cluster analysis. In data mining, cluster analysis approaches have shown to be a helpful and productive tool [8]. However, the majority of data is gathered in random formats and groups, which makes analysis challenging, particularly when the data pieces' properties are unclear. Cluster analysis is one area of data mining that focuses on effectively classifying unlabeled data. The intentional grouping of unlabeled data into groups based on similarities in the characteristics and attributes of data objects is called data clustering. The objective of cluster formation is to increase the similarity between data objects in a cluster compared to those in other clusters [9]. Data mining involves extracting extensive datasets to uncover underlying patterns. Within the field of data mining, various subjects include association rule mining, data clustering, and data classification [10]. Unsupervised learning places significant emphasis on the extensively researched field of clustering, as evidenced by numerous surveys and publications. Finding inherent patterns in a set of unlabeled data using clustering algorithms can be regarded as the pinnacle of unsupervised learning [11]. The objective is to organize a set of elements into cohesive clusters, grouping similar elements together while separating dissimilar ones into different clusters [12]. One of the data mining approach algorithms is K-Means (KM).

N data points are sorted into k clusters using the KM approach, which minimizes the sum of squared distances between each data point and its nearest cluster center (centroid). KM generates initial centroids by randomly selecting k data points, which are then refined in two main processes. During the assignment stage, each data point is assigned to the cluster of the nearest centroid. The average of all the points assigned to a cluster during the update stage is used to find the cluster's centroid. Together, these two procedures result in a single iteration of the KM algorithm [13]. KM clustering aims to separate a set of n data points into k groups by allocating each data point to the closest cluster center, which serves as the cluster's representative point [14].

Children's health and nutritional status serve as essential indicators of the broader public nutrition landscape. Malnutrition presents challenges not only for families but also burdens the nation as a whole [10]. The 2022 nutrition reports from the Sukodono community health center provided the data for this investigation. Numerous studies have shown that machine learning and the data mining approach can assist health authorities in processing public health data, particularly data on toddler stunting and malnutrition. This study cluster edits toddler data using a data mining technique and the KM method, specifically in the village of Jumput Rejo Sukodono, Sidoarjo.

2. METHOD

2.1. Research design

The Sukodono Community Health Center provided information on stunting in Jumput Rejo hamlet. The three primary steps in the data mining process are the application of a classification algorithm, preprocessing techniques, and an initial data overview. This study's data mining methodology maps instances of malnutrition in children under five using the KM clustering algorithm. Function optimization can be included into KM optimization to accurately identify the initial number of clusters. The elbow approach, which determines the ideal number of clusters by creating a “elbow” at a particular point, is used to develop and implement the model. The data is an anthropometric dataset of toddlers totaling 1,070. The dataset

contains a total of 6 attributes, which represent anthropometric data for toddlers. The attributes included in the dataset are as: name of toddler, sex, age, weight, height, and name of the integrated healthcare center. Before performing the clustering process on the toddlers' data, it is essential to conduct data preprocessing. Out of 1,070 data points, 583 complete data entries were obtained. A sample of the dataset is presented in Table 1.

Table 1. Toddler anthropometric dataset

No	Integrated health care center	Name	Sex	Age (month)	Weight (kg)	Height (kg)
1	Ciro1	Arbian Azza	1	28	11	87
2	Ciro1	M. Reza	1	59	16.4	110
…..	……..	…….	…..	…….	……	…….
1070	Surya Asri 2B	Nayla Nursalsabila	2	-1	0	0

In addition, there is a method for using name imputation to replace lost data. One method of imputation is known as correlation-based imputation [15]. The nullify the missing values before imputation (NMVI) imputation technique proposes splitting the data into whole and incomplete subsets. For each class with lacking data, an upper limit is set. This approach allows the model to more accurately estimate missing values and bring them closer to the actual values [16].

2.2. Data normalization

We can use data scaling to do normalize the data. Data scaling is a data normalization technique. Data normalization is the method of standardizing multiple variables to possess a uniform range of values, avoiding extremes that could complicate statistical analysis. Feature wise normalization is one of effective way to normalizing data [17], min-max normalization [18]. There are two categories of data scaling methods: the initial category employs metrics related to size like mean or median, while the second category utilizes measures of data spread (such as standard deviation or median absolute deviation) [19].

As a component of our data preprocessing, we perform data normalization, which involves scaling the original data values to fit within a designated narrow range of [0, 1] [20]. Data normalization is employed to scale an attribute's data to fit within a narrower range, such as -1 to 1 or 0 to 1. The normalization process as (1):

$$Normalized\ value = \frac{Initial\ value - Minimum\ value}{Max\ value - Minimum\ value} \quad (1)$$

The values of the variables will be normalized into the range 0-1.

2.3. Elbow method

Incorporating KM optimization into the objective function enables precise determination of the initial cluster count. This optimization uses the Elbow method, which identifies the ideal number of clusters by detecting the point where a distinct “elbow” appears.

An algorithm for calculating K for K-Means using the Elbow approach [21]:

1. Start
2. Set K to its initial value.
3. Raise the value of K.
4. For every K value, calculate the sum of squared error outcomes.
5. Analyze the sum of squared error findings to determine the value of K at which a notable decline occurs.
6. Set the K value that was determined.
7. End

The formula of sum of square error (SSE) is as following:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where n, y_i, and $\hat{y}_i$ are the number of data, the value of each point, and the average of all values, respectively.

2.4. K-Means

The following is an outline of the KM clustering procedure [10]:

a. Step 1: begin by defining k, which denotes the number of clusters
b. Step 2: provide an initial partition that divides the data into k clusters. This can be done by either randomly sampling the data or following a systematic approach: begin by assigning the first training sample to a single-element cluster. For the remaining samples (totaling N−k), each should be assigned to the cluster that has the closest centroid. After this stage is finished, recalculate the centroid for each newly formed cluster.
c. Step 3: determine the distance between each sample and the centroid of each cluster by evaluating each one separately. Reassign a sample to the closest cluster if it is allocated to one but not the closest one, updating the centroid and the sample's cluster membership in the process.
d. Step 4: continue Step 3 until you get the desired outcome, which is when every training sample is assigned correctly and no changes are needed. Every data point should be assigned to a different cluster as a centroid if the volume of data is fewer than the number of clusters. This way, each centroid will represent a different cluster. After calculating the distance between each data point and each centroidal, assign each data point to the cluster with the closest centroid when the data volume exceeds the number of clusters. If there is any doubt, recalculate the centroid's position using the information that is currently available, and then reassign all of the information to the updated centroid. Until no data points are sent to other clusters, keep doing this.

3. RESULTS AND DISCUSSION

3.1. Data processing

This research was conducted to classify stunting data on toddlers in an area in order to obtain an overview of which areas experience cases of babies suffering from stunting in the village of Jumput Rejo Sukodono, Sidoarjo. Data preprocessing is done by retrieve transaction data purposively sampling by cleaning the data incomplete. Data cleaning is performed on data that has a value of NaN. We also need to pay attention to the data type of our dataset. To carry out the clustering process, the dataset we have must be in the form of a numeric data type. Type of dataset are float64 and object. The column of dataset is no, name of healthcare center, name, sex, age, weight, and height. By using this formula, we will change the data type of our dataset to a numeric data type. Especially for the age, weight and height columns.

3.1.1. Eliminating outlier

The existence of this outlier data will cause deviations from the results of data analysis. An outlier is an observation whose point of observation deviates far from the data pattern. Therefore, the outlier data needs to be removed. Identifying outliers is a significant approach within the field of data mining due to its numerous important uses. It can be applied to eliminate unwanted noise from data or investigate particular data points that diverge significantly from their surrounding observations [22]. Numerous methods are employed for outlier removal. A novel method that maintains privacy while increasing clustering performance is provided by the outlier-eliminated differential privacy (OEDP) KM algorithm. Yu *et al.* [23], utilizing the two outlier identification methods mentioned above to determine each object's degree of outline [24], outlier detection simultaneously [25], to remove outliers adaptively, the traditional Tukey rule is altered [26], use the Z-Score [27], semi supervised outlier detection [28].
Z-Score formula:

$$Z\ Score = \frac{(x - \bar{x})}{\sigma}$$

where $\bar{x}$ is average data and σ is standard deviation.
An outlier is defined as a data point x for which the absolute values of its Z-Scores exceed 3.

From Figure 1, information is obtained, namely 3 data whose position is far from other data. These 3 data deviate from the data pattern and are outliers. With a certain process, it will be processed to eliminate outlier data. The process of eliminating outlier data using the Z-Score formula. A popular method for identifying outliers in datasets is the z-score test [22], [27]. From Figure 1, it is evident that only 3 data (0.6%) that deviates far from the data pattern. So that after removing the outlier data, the data is 580 toddler data.

3.1.2. Data normalization

Normalization of data involves adjusting multiple variables to share a consistent range of values, ensuring that none of them are excessively large or small, thus facilitating more straightforward statistical analysis. Data normalization is the procedure of ensuring that multiple variables share a consistent range of values, avoiding extremes that could complicate statistical analysis. We can use technique data scaling to do normalization the data.

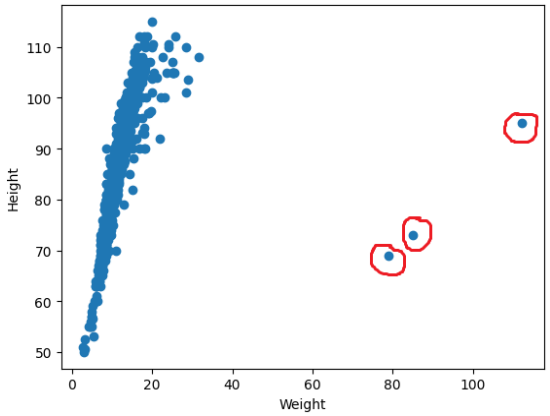

Figure 1. Outlier data visualizations

A fundamental and commonly used normalization technique is the min-max normalization method. This approach scales the lowest value in the dataset to 0 and the highest to 1, with all other values proportionally adjusted to fit within the range of 0 to 1 [29]. The lowest value of each feature is deducted from its matching values, and the result is then divided by the range of the feature (the maximum value minus the minimum value).

$$X_{New} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$$

To calculate the range of values, remove the minimum value of a feature from each of its values. This yields the highest value of the feature less its minimum value. Using this method, a new normalized value between 0 and 1 is produced. In this instance, the two variables utilized for data normalization are height and age.

3.2. Process with clustering algorithm elbow K-Means

3.2.1. Optimization of K-Means with the elbow method

Prior to using the KM method, it is essential to determine how many clusters there are in the data. The number of clusters will determine the final result of the data grouping process using the KM algorithm. Figure 2 will affect how the data is grouped by the KM method. The elbow technique is applied in this work to ascertain the ideal number of clusters to build. Using the elbow method to calculate the value of K approach can be used to calculate with metric within cluster sum of square (WCSS). The smaller the WCSS score, the better. Using the method described in section 2.3 for data after normalization, it is found that the significant change in the WCSS value occurs at the value K=4 so that it is obtained that the optimum number of clusters is 4 clusters. An algorithm for calculating K for K-Means using the Elbow approach [21]. The application of KM involves organizing data into groups based on a specified number of clusters, in this case, there are 4 groups. Determining the appropriate value of K is crucial, and the elbow method is employed for this purpose, utilizing the WCSS metric. A lower WCSS score indicates better clustering. The graph clearly shows a notable change in the WCSS value at K=4, indicating that the number of clusters that is ideal is 4.

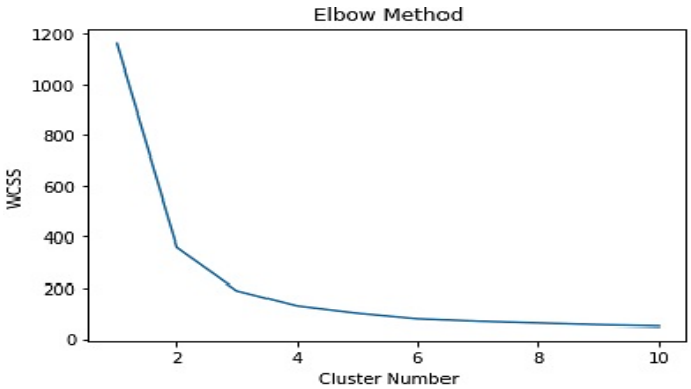

Figure 2. Elbow method chart

In other research, elbow method uses to cluster for document. The elbow method will be grouped on the dataset by its range by submitting documents in all three domains of the two datasets [30]. This study employs the Elbow method, a cluster analysis technique, to make an informed decision regarding the number of clusters, considering the comparative values (derived from SSE calculations for each cluster) where an 'elbow' point is formed [31].

In Table 2, we can infer from the data that the number of clusters K=4 has the lowest SSE value. The value is 3756166.81. Hasugian *et al.* [31] research demonstrates this as well. A thorough analysis of earlier studies on toddler clustering has been made possible by the testing of the KM method and the Elbow Method. A considerable proportion of children, especially those in cluster 4, have stunted growth, according to the standard anthropometric table used to assess children's nutritional condition.

Table 2. Number of cluster (K) and value of SSE

No	Number of cluster (K)	Value of SSE
1.	2	14135848.55
2.	3	6250691.73
3.	4	3756166.81

3.2.2. K-Means clustering result

The KM clustering technique yields four clusters from the elbow method. The KM technique was then used to carry out the clustering process, as previously mentioned in section 2.4. Table 3 displays the outcome of the data clustering process.

Table 3. Number of value toddlers in cluster

Integrated health care center	Toddlers value Cluster 1	Toddlers value Cluster 2	Toddlers value Cluster 3	Toddlers value Cluster 4
Ciro 1	18	24	13	20
Ciro 2	17	15	23	22
Citra Surya Mas	10	4	10	12
Jumput Wetan	17	7	12	12
Jumput Rejo Indah	12	11	20	22
Keling	17	20	18	28
Puri Sejahtera 3	5	5	6	6
Surya Asri 2b	12	9	13	8
Citra Gading	24	11	19	12
Kedung 1	18	18	16	14
Sum of cluster member	150	124	150	156

From the result, it is found that Cluster 1 has a total of 150 cluster members of consisting of integrated healthcare center group. where Cluster 1 has a relatively smaller range of values for age and height than the other clusters. The number of Cluster 1 data for the male sex is 75 data, while for the female sex there are 75 data. Cluster 2 total data 124 member of consisting of integrated healthcare center group. Cluster 2 has a total of 124 cluster members, where Cluster 2 has a range of values for age and height which is relatively higher than Cluster 1 and Cluster 3. The number of Cluster 2 data for the male sex is 73 data, while for the female sex there are 51 data. Cluster 3 total data 125 member of consisting of integrated healthcare center group. Cluster 3 has a total of 150 cluster members, where Cluster 3 has a range of values for age and height that are somewhat lower than cluster 1 and somewhat higher than Cluster 2 and Cluster 4. The number of cluster 3 data for the male sex is 74 data, while for the female sex there are 76 data. Cluster 4 total data 156 member of consisting of integrated healthcare center group. Cluster 4 has a total of 156 cluster members, where Cluster 4 has a range of values for age and height that are relatively higher than the other clusters. The number of Cluster 4 data for the male sex is 82 data, while for the female sex there are 74 data.

The results of visualization of grouping data for toddlers is presented in Figure 2. Cluster 4 has the highest age and height among other toddler anthropometric data clusters. In cluster 3 has a range of values for age and height they are comparatively lower than Clusters 2 and 4 and higher than Cluster 1. In Cluster 2 has a range of values for age and height which is relatively higher than Cluster 1 and Cluster 3. From the Figure 3, we also know that Cluster 1 has a relatively smaller range of values for age and height than the other clusters. If refer to the standard anthropometric table for assessing children's nutritional status, it is found that many children are stunted especially toddlers who are members of Cluster 4.

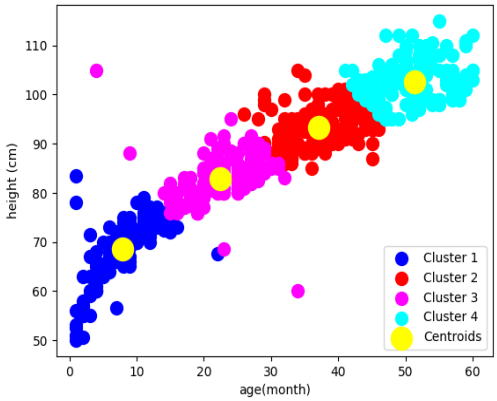

Figure 3. Visualization of the results of clustering of toddler anthropometric data

4. CONCLUSION

Our study focuses on clustering anthropometric data of toddlers. A total of 580 children's anthropometric measurements were identified for grouping based on their similarity. The elbow approach is used to optimize the number of clusters in the data grouping process, which uses the KM algorithm. Four clusters were generated from the cluster optimization findings, and these will be utilized as input to determine which clusters were formed during the data grouping process using the KM technique. Especially in Cluster 4 has a total of 156 cluster members, where Cluster 4 has a range of values for age and height that are relatively higher than the other clusters. The number of Cluster 4 data the male sex is 82 data, while for the female sex there are 74 data. New information will be generated from the toddler clustering process in the village of Jumput Rejo Sukodono Sidoarjo in terms of age and height that many children are stunted especially toddlers who are members of Cluster 4 based on standard anthropometric table for assessing children's nutritional status.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amir Ali	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Purwanto		✓			✓		✓	✓		✓		✓		
Mundakir		✓			✓		✓	✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author [AA] upon reasonable request.

REFERENCES

[1] D. S. H. Putra, I. G. Wiryawan, E. R. Pristiwaningsih, E. Mulyadi, P. Destarianto, and K. Agustianto, "Development of Malnutrition Early Detection Application in Toddlers based on Geographic Information System," in *Proceedings of the 2nd International Conference on Social Science, Humanity and Public Health (ICOSHIP)*, 2022, vol. 645, pp. 175-181, doi: 10.2991/assehr.k.220207.028.

[2] F. H. Bitew, C. S. Sparks, and S. H. Nyarko, "Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia," *Public Health Nutrition*, vol. 25, no. 2, pp. 269–280, Oct. 2022, doi: 10.1017/S1368980021004262.

[3] A. S. Shamsuddin, W. A. M. A. Bakar, S. N. S. Ismail, N. H. Jaafar, W. M. Yassin, and M. Norhizat, "A Review of Spatial Analysis Application in Childhood Malnutrition Studies," *Malaysian Journal of Medical Sciences*, vol. 29, no. 5, pp. 24–38, Oct. 2022, doi: 10.21315/mjms2022.29.5.4.

[4] M. d. Onis and F. Branca, "Childhood stunting: A global perspective," *Maternal and Child Nutrition*, vol. 12, pp. 12–26, May 2016, doi: 10.1111/mcn.12231.

[5] C. P. Stewart, L. Iannotti, K. G. Dewey, K. F. Michaelsen, and A. W. Onyango, "Contextualising complementary feeding in a broader framework for stunting prevention," *Maternal and Child Nutrition*, vol. 9, pp. 27–45, Sep. 2013, doi: 10.1111/mcn.12088.

[6] L. Yin *et al.*, "Nutritional features-based clustering analysis as a feasible approach for early identification of malnutrition in patients with cancer," *European Journal of Clinical Nutrition*, vol. 75, no. 8, pp. 1291–1301, Aug. 2021, doi: 10.1038/s41430-020-00844-8.

[7] R. E. Black and R. Heidkamp, "Causes of stunting and preventive dietary interventions in pregnancy and early childhood," in *Nestle Nutrition Institute Workshop Series*, vol. 89, pp. 105–113, 2018, doi: 10.1159/000486496.

[8] R. Mussabayev, N. Mladenovic, B. Jarboui, and R. Mussabayev, "How to Use K-means for Big Data Clustering?," *Pattern Recognition*, vol. 137, p. 109269, May 2023, doi: 10.1016/j.patcog.2022.109269.

[9] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[10] S. Winiarti, H. Yuliansyah, and A. A. Purnama, "Identification of Toddlers' nutritional status using data mining approach," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 164–169, 2018, doi: 10.14569/IJACSA.2018.090122.

[11] A. Aradnia, M. A. Haeri, and M. M. Ebadzadeh, "Adaptive Explicit Kernel Minkowski Weighted K-means," *Information Sciences*, vol. 584, pp. 503–518, Jan. 2022, doi: 10.1016/j.ins.2021.10.048.

[12] P. Mansueto and F. Schoen, "Memetic differential evolution methods for clustering problems," *Pattern Recognition*, vol. 114, p. 107849, Jun. 2021, doi: 10.1016/j.patcog.2021.107849.

[13] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognition*, vol. 93, pp. 95–112, Sep. 2019, doi: 10.1016/j.patcog.2019.04.014.

[14] H. Hu, J. Liu, X. Zhang, and M. Fang, "An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis," *Pattern Recognition*, vol. 139, p. 109404, Jul. 2023, doi: 10.1016/j.patcog.2023.109404.

[15] I. Curioso *et al.*, "Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-based Imputation," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 6, p. 101562, Jun. 2023, doi: 10.1016/j.jksuci.2023.101562.

[16] H. V. Bhagat and M. Singh, "NMVI: A data-splitting based imputation technique for distinct types of missing data," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, p. 104518, Apr. 2022, doi: 10.1016/j.chemolab.2022.104518.

[17] D. Singh and B. Singh, "Feature wise normalization: An effective way of normalizing data," *Pattern Recognition*, vol. 122, p. 108307, Feb. 2022, doi: 10.1016/j.patcog.2021.108307.

[18] P. J. Jones *et al.*, "FilterK: A new outlier detection method for k-means clustering of physical activity," *Journal of Biomedical Informatics*, vol. 104, p. 103397, Apr. 2020, doi: 10.1016/j.jbi.2020.103397.

[19] J. Walach, P. Filzmoser, and K. Hron, "Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences," in *Comprehensive Analytical Chemistry*, vol. 82, pp. 165–196, 2018, doi: 10.1016/bs.coac.2018.06.004.

[20] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, 2019, doi: 10.1016/j.imu.2019.100179.

[21] R. Nainggolan, R. P. Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *Journal of Physics: Conference Series*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[22] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of spatial outlier by using improved Z-score test," in *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019-April, pp. 788–790, Apr. 2019, doi: 10.1109/icoei.2019.8862582.

[23] Q. Yu, Y. Luo, C. Chen, and X. Ding, "Outlier-eliminated k-means clustering algorithm based on differential privacy preservation," *Applied Intelligence*, vol. 45, no. 4, pp. 1179–1191, Dec. 2016, doi: 10.1007/s10489-016-0813-z.

[24] F. Jiang, G. Liu, J. Du, and Y. Sui, "Initialization of K-modes clustering using outlier detection techniques," *Information Sciences*, vol. 332, pp. 167–183, Mar. 2016, doi: 10.1016/j.ins.2015.11.005.

[25] G. Gan and M. K. P. Ng, "K-Means Clustering With Outlier Removal," *Pattern Recognition Letters*, vol. 90, pp. 8–14, Apr. 2017, doi: 10.1016/j.patrec.2017.03.008.

[26] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6365–6376, Sep. 2022, doi: 10.1016/j.jksuci.2021.07.003.

[27] E. J. Jamshidi, Y. Yusup, J. S. Kayode, and M. A. Kamaruddin, "Detecting outliers in a univariate time series dataset using unsupervised combined statistical methods: A case study on surface water temperature," *Ecological Informatics*, vol. 69, p. 101672, Jul. 2022, doi: 10.1016/j.ecoinf.2022.101672.

[28] Y. Zhao, H. Li, X. Yu, N. Ma, T. Yang, and J. Zhou, "An independent central point OPTICS clustering algorithm for semi-supervised outlier detection of continuous glucose measurements," *Biomedical Signal Processing and Control*, vol. 71, p. 103196, Jan. 2022, doi: 10.1016/j.bspc.2021.103196.

[29] S. I. Kim, Y. Noh, Y. J. Kang, S. Park, J. W. Lee, and S. W. Chin, "Hybrid data-scaling method for fault classification of compressors," *Measurement: Journal of the International Measurement Confederation*, vol. 201, p. 111619, Sep. 2022, doi: 10.1016/j.measurement.2022.111619.

[30] P. Mekkamol and C. Jareanpon, "the Development of a New Hybrid K-Means and Elbow Method (C-Algorithm) for Multiple

Domain Clustering," *ICIC Express Letters*, vol. 17, no. 3, pp. 269–278, 2023, doi: 10.24507/icicel.17.03.269.
[31] P. M. Hasugian, B. Sinaga, J. Manurung, and S. A. Al Hashim, "Best Cluster Optimization with Combination of K-Means Algorithm and Elbow Method Towards Rice Production Status Determination," *International Journal of Artificial Intelligence Research*, vol. 5, no. 1, Jun. 2021, doi: 10.29099/ijair.v6i1.232.

BIOGRAPHIES OF AUTHORS

Amir Ali received the Engineer degree in Informatic Engineering from 17 Agustus (UNTAG) Surabaya University in 2008, and received Master of Information Technology in Sekolah Tinggi Teknik Surabaya, Indonesia, in 2016. Currently he is lecturer in Sekolah Tinggi Ilmu Kesehatan Yayasan RS Dr. Soetomo in Surabaya and research student in Diponegoro University in 2022. His research interests include data mining, machine learning, web programming, and human computer interaction. He can be contacted at email: amir.consulting@gmail.com.

Purwanto is a Professor at the Department of Information System, Diponegoro University, Indonesia, where he has been a lecturer at Diponegoro University. He graduated with a first-class honors Engineer degree in Chemical Engineering from Diponegoro University, Indonesia, in 1985, and an D.E.A. in Computer Engineering from Institut National Polytechnique De Toulouse University, France in 1991. Completed his Ph.D. in Computer Engineering from Institute National Polytechnique De Toulouse, in 1994. His research interests are primarily in the area of Computer Engineering where he is the author/co-author of over 70 research publications. He can be contacted at email: purwanto.profundip@gmail.com.

Mundakir received Bachelor of nursing from Airlangga University, Indonesia, Master of Nursing from Indonesia University, Indonesia, and Ph.D. in health Sciences, Airlangga University, Indonesia. He was vice chancellor Muhammadiyah Surabaya University. Currently, he is a lecturer at Faculty of Health Sciences, Muhammadiyah Surabaya University, Surabaya, Indonesia. His research interests are nursing science and health information system management. He can be contacted at email: mundakir@um-surabaya.ac.id.